



ISSN 2071-2898 (Print)  
ISSN 2071-2901 (Online)

**Бахвалов П.А.**

О точности разрывного  
метода Галёркина для  
одномерного уравнения  
переноса при длительном  
счёте

**Рекомендуемая форма библиографической ссылки:** Бахвалов П.А. О точности разрывного метода Галёркина для одномерного уравнения переноса при длительном счёте // Препринты ИПМ им. М.В.Келдыша. 2017. № 134. 24 с. doi:[10.20948/prepr-2017-134](https://doi.org/10.20948/prepr-2017-134)  
URL: <http://library.keldysh.ru/preprint.asp?id=2017-134>

**Ордена Ленина  
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ  
имени М.В.КЕЛДЫША  
Российской академии наук**

**П. А. Бахвалов**

**О точности разрывного метода Галёркина  
для одномерного уравнения переноса  
при длительном счёте**

**Москва — 2017**

## Бахвалов П. А.

О точности разрывного метода Галёркина для одномерного уравнения переноса при длительном счёте

Метод нестационарного корректора применяется к исследованию точности метода Галёркина с разрывными базисными функциями для одномерного уравнения переноса с постоянной скоростью. Начальные данные предполагаются достаточно гладкими. Доказывается, что для произвольной неравномерной сетки численная ошибка удовлетворяет оценке  $\|e\| = O(h^{k+1} + th^{2k+1})$ , где  $k$  – порядок используемого полинома.

**Ключевые слова:** разрывный метод Галёркина, аппроксимация и точность, суперсходимость

## Pavel Alexeevich Bakhvalov

On the long-time simulation accuracy of the discontinuous Galerkin method for 1D transport equation

Unsteady corrector method is applied to study the DG method for 1D transport equation with constant velocity. Only smooth solutions are considered. On arbitrary non-uniform mesh we prove that the numerical error satisfies  $\|e\| = O(h^{k+1} + th^{2k+1})$  where  $k + 1$  is the number of DOFs at one cell.

**Key words:** discontinuous Galerkin method, consistency and accuracy, superconvergence, long-time simulation accuracy

## Оглавление

Введение . . . . .	3
Разрывный метод Галёркина . . . . .	4
О методе нестационарного корректора . . . . .	6
Алгоритм исследования схемы общего вида . . . . .	8
Примеры . . . . .	12
Квадратуры типа Гаусса . . . . .	14
Метод Галёркина с точками коллокации в узлах квадратуры . . . . .	15
Случай неравномерной сетки . . . . .	19
Случай произвольного базиса . . . . .	20
Тестирование . . . . .	21
Заключение . . . . .	22
Список литературы . . . . .	23

## Введение

Метод Галёркина с разрывными базисными функциями (discontinuous Galerkin method, DG) был предложен в [1] и в настоящее время является одной из наиболее распространённых схем для решения гиперболических систем уравнений на неструктурированных сетках. В этом методе сеточная функция представляет собой многочлен некоторого порядка  $p$  в пределах сеточного элемента, причём многочлены на каждом сеточном элементе независимы друг от друга.

Всюду в настоящей работе будем рассматривать уравнение переноса с постоянной скоростью и предполагать, что точное решение достаточно гладкое. Также будем считать, что оператор проектирования решения на пространство сеточных функций точен на полиномах порядка  $p$  (например, таковой является  $L_2$ -проекция). Тогда в многомерном случае на неструктурированной сетке при сохранении качества сеточных элементов DG обладает аппроксимационной ошибкой  $O(h^p)$  и позволяет найти решение с точностью  $O(h^{p+1/2})$  в норме  $L_2$  [2]. В одномерном случае на неравномерной сетке порядок аппроксимации также равен  $p$ , но порядок точности равен  $p + 1$  [3].

На равномерных (трансляционно-симметрических) сетках порядок аппроксимации в общем случае равен  $p$ . Порядок точности как в одномерном, так и в многомерном случае равен  $p + 1$ . При этом слагаемое порядка  $h^{p+1}$  в ошибке решения ограничено величиной, не зависящей от времени, а старший член в ошибке решения, неограниченно растущий со временем, имеет порядок  $h^{2p+1}$  [4]. Фактически этот результат следует из  $(2p + 1)$ -го порядка точности разрывного метода Галёркина в смысле пост-обработки численного решения, доказанной в одномерном случае в [5].

Оценки точности метода DG были получены и для более сложных задач. Случаи нелинейных уравнений и систем уравнений исследованы в [6–8]. Пост-обработка численного решения [5] была обобщена на случай линейных уравнений с переменными коэффициентами и нелинейных уравнений, а также задач с граничными условиями в [9–11].

В [12] было проведено сравнение разрывного метода Галёркина с кусочно-линейными базисными функциями и метода спектральных объёмов (Spectral Volume) на основе полиномов 1-го порядка. Рассматривался случай одномерного уравнения переноса с постоянным коэффициентом на равномерной сетке. Сравнение проводилось при помощи Фурье-анализа. Для этого в обоих методах в пространстве полиномов на каждой ячейке был выбран коллокационный базис, что позволило представить эти схемы как (неоднородные) конечно-разностные схемы на равномерных сетках.

В настоящей работе проводится исследование точности метода DG на неравномерной одномерной сетке при помощи метода нестационарного кор-

ректора [13]. Для этого, как и в [12], используется конечно-разностное представление схемы DG, основанное на коллокационном базисе. Приводится новое доказательство того, что старший член в ошибке решения, неограниченно растущий со временем, имеет порядок  $h^{2p+1}$ .

## Разрывный метод Галёркина

В настоящей работе будем рассматривать начальную задачу для одномерного уравнения переноса с единичной скоростью:

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial x} = 0, \quad t > 0, \quad x \in \mathbb{R}. \quad (1)$$

Пусть начальные данные заданы функцией  $u(0, x) = u_0(x)$ , в каждой точке имеющей  $2p + 1$  липшицевых производных и ограниченной при  $x \rightarrow \pm\infty$  вместе с этими производными. Пусть на числовой оси заданы узлы  $X_i, i \in \mathbb{N}$ , и введём  $h_i = X_{i+1} - X_i$ .

На модельном отрезке  $(0,1)$  введём набор линейно независимых базисных функций  $\{\phi_j(x)\}, j = 0, \dots, p$ , каждая из которых является полиномом порядка  $p$ . Определим матрицы масс и градиентов  $M$  и  $G$  с коэффициентами

$$m_{jk} = \int_0^1 \phi_j(x)\phi_k(x)dx, \quad g_{jk} = \int_0^1 \phi'_j(x)\phi_k(x)dx.$$

На числовой прямой введём базисные функции  $\{\psi_{i,j}(x)\}$  следующим образом:

$$\psi_{i,j}(x) = \begin{cases} \phi_j((x - X_i)/h_i), & X_i < x < X_{i+1}; \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Домножим (1) на базисную функцию  $\psi_{i,j}(x)$  и проинтегрируем по  $x$ . Заменяя  $x$  на  $X_i + \xi h_i$  и применяя интегрирование по частям во втором слагаемом, получаем

$$\begin{aligned} & \int_0^1 \phi_j(\xi) \frac{\partial u}{\partial t}(t, X_i + \xi h_i) d\xi + \phi_j(1)u(t, X_{i+1}) - \\ & - \phi_j(0)u(t, X_i) - \int_0^1 \phi'_j(\xi)u(t, X_i + \xi h_i) d\xi = 0. \end{aligned} \quad (3)$$

Обозначим набор неизвестных через  $u_{i,j}(t)$ , где  $i \in \mathbb{N}$  – номер ячейки, а  $j = 0, \dots, p$  – индекс степени свободы на ячейке. При выбранном базисе

метод Галёркина с разрывными базисными функциями для этой задачи записывается в виде системы ОДУ

$$\begin{aligned} h_i \sum_{k=0}^p m_{jk} \frac{du_{i,k}}{dt} + \phi_j(1) \sum_{k=0}^p \phi_k(1) u_{i,k} - \\ - \phi_j(0) \sum_{k=0}^p \phi_k(1) u_{i-1,k} - \sum_{k=0}^p g_{jk} u_{i,k} = 0. \end{aligned} \quad (4)$$

Под  $\phi_k(0)$  и  $\phi_k(1)$  понимаются предельные значения изнутри отрезка  $(0, 1)$ .

Выбор базиса в пространстве многочленов на каждой ячейке влияет только на представление решения, но не меняет разностную схему по существу. Чтобы воспользоваться методом нестационарного корректора для исследования точности этой схемы, необходимо использовать коллокационный базис. Он строится следующим образом. Выберем на отрезке  $[0, 1]$  некоторым образом точки  $x_j$ ,  $j = 0, \dots, p$ , причём  $x_j \neq x_k$  при  $j \neq k$ . Положим, что  $\{\phi_j(x)\}$  – полиномы порядка  $p$ , удовлетворяющие условию  $\phi_j(x_k) = \delta_{jk}$ , где  $\delta_{jk}$  – символ Кронекера. После этого базисные функции определим формулой (2).

Начальные данные могут проектироваться на сетку различным образом. Наиболее часто используется ортогональная проекция: на каждой ячейке  $(X_i, X_{i+1})$  начальное значение сеточной функции определяется из решения системы уравнений

$$\int_0^1 \left( \sum_{m=0}^p \phi_m(x) u_{i,m}(0) \right) x^k dx = \int_0^1 u_0(X_i + x h_i) x^k dx.$$

В случае коллокационного базиса начальные данные удобно задавать поточечно:

$$u_{i,j}(0) = u_0(X_i + x_j h_i). \quad (5)$$

Мы будем использовать способ (5), хотя метод нестационарного корректора позволяет работать и с интегральным оператором проектирования.

**Утверждение 1.** *Если решение  $u(t, x) = u_0(x - t)$  является многочленом порядка не выше  $p$ , то схема (4) является точной.*

Для доказательства достаточно заметить, что в этом случае решение точно разлагается по базисным функциям:  $u(t, x) = \sum_i \sum_j \psi_{i,j}(x) u_{i,j}(t)$ , и равенство (4) вырождается в верное равенство (3).

Перепишем метод Галёркина в матричном виде. Для этого введём обозначение  $\mathbf{u}_i = \{u_{i,j}\}, j = 0, \dots, p$ . Имеем

$$h_i M \frac{d\mathbf{u}_i}{dt} + A\mathbf{u}_i + B\mathbf{u}_{i-1} = 0, \quad (6)$$

где  $A$  и  $B$  – матрицы  $(p+1) \times (p+1)$  с коэффициентами

$$a_{jk} = \phi_j(1)\phi_k(1) - g_{jk} = \phi_j(1)\phi_k(1) - \int_0^1 \phi_j'(x)\phi_k(x)dx, \quad (7)$$

$$b_{jk} = -\phi_j(0)\phi_k(1).$$

Опишем алгоритм исследования схемы общего вида (6) при  $h = const$  методом нестационарного корректора. Позже для рассмотрения разрывного метода Галёркина произвольного порядка мы подставим конкретные выражения для матриц  $M$ ,  $A$  и  $B$ , а затем перейдём к случаю неравномерной сетки.

## О методе нестационарного корректора

Далее вплоть до параграфа, посвящённого анализу разрывного метода Галёркина на неравномерной сетке, будем рассматривать случай равномерной сетки с шагом  $h$  и узлами  $X_i = ih$ .

Для исследования точности схемы (6) на равномерной сетке применим метод нестационарного корректора [13] в условиях наличия периода в разностной задаче. Введём обозначения, соответствующие [13]. Сеточная функция  $u = \{u_j\}, j = [\eta, \xi] \in M, \eta \in \mathbb{Z}^{\tilde{d}}$  – индекс блока,  $\xi \in M^0$  – индекс неизвестной внутри блока,  $M = M^0 \times \mathbb{Z}^{\tilde{d}}$  – множество индексов. В нашем случае, поскольку мы рассматриваем одномерную задачу с периодичностью разностной схемы по одному направлению,  $\tilde{d} = 1$ .  $\eta$  – индекс блока – совпадает с номером ячейки:  $\eta \equiv i$ .  $M^0$  есть множество степеней свободы на одной ячейке:  $M^0 = \{0, \dots, p\}$ .

Метод нестационарного корректора был разработан автором для разностных схем вида

$$\frac{du_j}{dt} + \sum_{k \in M} L_{jk} u_k = 0. \quad (8)$$

Схема (6) при  $h = const$  представима в этом виде, если положить

$$L_{[\eta, \xi], [\eta', \xi']} = \frac{1}{h} \times \begin{cases} (M^{-1}A)_{\xi, \xi'}, & \eta' = \eta; \\ (M^{-1}B)_{\xi, \xi'}, & \eta' = \eta - 1; \\ 0, & otherwise. \end{cases} \quad (9)$$

Очевидно, что разностный оператор удовлетворяет условию периодичности:

$$L_{[\eta+n, \xi], [\eta'+n, \xi']} = L_{[\eta, \xi], [\eta', \xi']}, \quad \forall n \in \mathbb{Z}. \quad (10)$$

Введём матрицу

$$\hat{L} = \{\hat{L}_{jk}\}, \quad \hat{L}_{jk} = \sum_{\eta \in \mathbb{Z}} L_{[\eta^0, j], [\eta^0 + \eta, k]}.$$

Ввиду (10) эта матрица не зависит от  $\eta^0$ . В нашем случае она равна

$$\hat{L} = \frac{1}{h} M^{-1} (A + B). \quad (11)$$

Обозначим через  $\Pi$  оператор проектирования, используемый для задания начальных данных и проектирования на сетку точного решения. Формула (5) подразумевает, что он определяется формулой

$$(\Pi u)_{[\eta, \xi]} = u((\eta + x_\xi)h). \quad (12)$$

Видно, что при пространственной трансляции функции  $u$  на величину  $h$  её проекция смещается на один блок:

$$(\Pi T_h u)_{[\eta, \xi]} = (\Pi u)_{[\eta+1, \xi]}, \quad (T_h u)(x) = u(x + h). \quad (13)$$

Условия (10) и (13) гарантируют, что свойства схемы одинаковы во всех сеточных блоках, благодаря чему задача их анализа сводится к конечномерной. В частности, нестационарные корректоры всех порядков одинаковы во всех сеточных блоках.

Определение нестационарных корректоров, введённое в [13], достаточно громоздкое, и мы его приводить не будем. В отличие от [13], нам будет достаточно рассмотрения стационарных операторов проектирования. Если уравнение (6) решается явным методом Рунге – Кутты достаточного порядка точности, причём шаг по времени находится строго внутри его области устойчивости, то нестационарные корректоры имеют предельные значения при  $t \rightarrow \infty$ , совпадающие с коэффициентами оператора проектирования  $\tilde{\Pi}$ , которые будут рассматриваться в настоящей работе.

Пусть  $\tilde{\Pi}$  – некоторый оператор, сопоставляющий гладкой функции  $f : \mathbb{R} \rightarrow \mathbb{R}$  вектор с компонентами  $(\tilde{\Pi} f)_j$ ,  $j \in M$ . Будем называть ошибкой аппроксимации схемы (8) на функции  $f(x - t)$  в смысле  $\tilde{\Pi}$  значение выражения

$$\epsilon_j(f) = - \left( \tilde{\Pi} \frac{\partial f}{\partial x} \right)_j + \sum_{k \in M} L_{jk} \left( \tilde{\Pi} f \right)_k. \quad (14)$$

Пусть разностная схема точна на полиномах порядка  $\tilde{p} - 1$  в смысле  $\tilde{\Pi}$ , удовлетворяющего условию (13). Обозначим через  $f^{\tilde{p}}$  ошибку аппроксимации



на нормированном полиноме следующего порядка:

$$f_j^{\tilde{p}} = \epsilon_j \left( \frac{(x-t)^{\tilde{p}}}{\tilde{p}!} \right).$$

Очевидно,  $f_j^{\tilde{p}}$  не зависит от  $t$  и пространственной трансляции полинома. Поскольку в определение (9) оператора  $L$  входит матрица  $M^{-1}$ , удобно работать с матрицей  $hM\hat{L} = A + B$  и домноженными на матрицу  $M$  векторами аппроксимационной ошибки  $Mf^{\tilde{p}}$ .

### Алгоритм исследования схемы общего вида

Рассмотрим схему общего вида (6) и предположим, что оператор проектирования  $\Pi$  имеет вид (12). Определим алгоритм (т. е. конечную процедуру), позволяющий определить формальный порядок точности и порядок точности при длинном времени счёта рассматриваемой разностной схемы.

Шаг 0. Вычислим матрицу  $hM\hat{L} = A + B$ . Убедимся, что схема точна на константе, то есть сумма в каждой строке матрицы  $hM\hat{L}$  равна 0. В этом случае матрица  $hM\hat{L}$  заведомо вырождена. Если условие точности на константе нарушается, сходимости численного решения к точному нет.

Шаг 1. Положим  $q = 0$ .

Шаг 2. Вычислим

$$\begin{aligned} \frac{1}{h^q} (Mf)_j^{q+1} &= \sum_{k=0}^p m_{jk} \left( \frac{x_k^q}{q!} + \sum_{i=1}^q C_k^{q+1-i} \frac{x_k^{i-1}}{(i-1)!} \right) - \\ &- \sum_{k=0}^p a_{jk} \left( \frac{x_k^{q+1}}{(q+1)!} + \sum_{i=1}^q C_k^{q+1-i} \frac{x_k^i}{i!} \right) - \\ &- \sum_{k=0}^p b_{jk} \left( \frac{(x_k-1)^{q+1}}{(q+1)!} + \sum_{i=1}^q C_k^{q+1-i} \frac{(x_k-1)^i}{i!} \right). \end{aligned} \quad (15)$$

Минимальное значение  $q + 1$ , для которого  $(Mf)^{q+1} \neq 0$ , будем называть порядком главного корректора  $\tilde{p}$ .

Значение  $f^{q+1}$  имеет смысл ошибки аппроксимации схемы (8), (9) на функции  $-(x-t-ih)^{q+1}/((q+1)!)$  в смысле оператора проектирования  $\tilde{\Pi}^q$  (17), который будет введён ниже.

Шаг 3. Составим систему уравнений

$$(hM\hat{L})C^{q+1} = \frac{1}{h^q} (Mf)^{q+1} \quad (16)$$

и проверим совместность этой системы. Если система несовместна, то порядок  $Q$  схемы при длинном времени счёта равен  $q$ . Формальный порядок точности равен  $P = \min\{Q, \tilde{p}\}$ . Исследование закончено.

Шаг 4. Если система совместна, то находим из неё решение  $C^{q+1}$ , ортогональное левым собственным векторам матрицы  $hM\hat{L}$ , соответствующим нулевым собственным значениям. Очевидно,  $C^{q+1}$  не зависит от  $h$ .

Шаг 5. Увеличиваем значение  $q$  на единицу и переходим к шагу 2.

Результатом работы алгоритма являются значения  $\tilde{p}$ ,  $P$  и  $Q$ . При этом справедлива альтернатива: либо  $Q = P = \tilde{p} - 1$ , либо  $Q \geq P = \tilde{p}$ . В последнем случае в результате работы алгоритма также находится набор коэффициентов  $C^q$ ,  $P \leq q \leq Q$ .

Приведённый алгоритм обобщается естественным образом на случай, если в разностной схеме (6) шаблон для определения производной  $du_i/dt$  включает в себя не только значения  $u_i$  и  $u_{i-1}$ , но и некоторые другие  $u_k$ ,  $k = -m, \dots, m$ ,  $m < \infty$ . В многомерном случае на трансляционно-симметрических сетках он обобщается на схемы вида

$$hM \frac{du_i}{dt} + Au_i + \sum_{\omega \in \Omega} B^{(\omega)} u_{i+\omega} = 0,$$

где  $i$  – мультииндекс ячейки, а  $\Omega$  – некоторое множество соседних ячеек, значения в которых используются для определения эволюции в данной ячейке.

**Замечание 1.** Условие ортогональности  $C^{q+1}$  левым собственным векторам  $hM\hat{L}$ , соответствующим нулевым собственным значениям, не является обязательным. Из всего множества решений оно отбирает то, которое является предельным значением нестационарного корректора порядка  $q + 1$  при  $t \rightarrow \infty$ , делённым на  $h^{q+1}$ . Прямой подстановкой можно убедиться, что выбор любого другого решения системы не влияет ни на параметры  $\tilde{p}$ ,  $P$  и  $Q$ , ни даже на следующие коэффициенты  $C$ , которые будут найдены в результате работы алгоритма.

**Утверждение 2.** Рассмотрим оператор проектирования  $\tilde{\Pi}^Q$ , определяемый равенством

$$(\tilde{\Pi}^Q u)_{i,j}^n = u(t^n, (i + x_j)h) + \sum_{q=\tilde{p}}^Q h^q C_j^q \frac{\partial^q u}{\partial x^q}(t^n, (i + x_j)h), \quad (17)$$

где число  $Q$  и коэффициенты  $C_j^q$  найдены в ходе выполнения описанного алгоритма, и  $Q \leq Q$ . Тогда в смысле оператора проектирования  $\tilde{\Pi}^Q$  схема точна на полиномиальных порядка  $Q$  решениях уравнения (1).

Доказательство. При  $Q = 0$  утверждение очевидно. Предположим, что оно выполнено для некоторого  $Q < Q$ . Тогда аппроксимационная ошибка на всех полиномиальных порядка  $(Q + 1)$  решениях уравнения (1) определяется только старшим членом этого полинома и пропорциональна ему. Кроме того, она не зависит от времени. Поэтому достаточно доказать для функции  $f(t, x) = (x - ih - t)^{Q+1}/(Q + 1)!$  при  $t = 0$ . Подставим функцию  $f(t, x)$  и определение проектора  $\tilde{\Pi}^{Q+1}$  (17) в формулу для аппроксимационной ошибки (14). Приравнявая полученное к нулю, получим систему (16), которая выполняется в силу алгоритма.

**Утверждение 3.** Пусть полудискретная схема (6) интегрируется во времени с ошибкой не выше  $O(\tau^Q)$ , причём  $\tau \leq \tilde{c}h$ . Пусть получающаяся при этом полностью дискретная схема является устойчивой с константой  $K$  в некоторой норме  $l_p$  (параметр нормы  $p$  не связан с порядком полиномов) или  $l_\infty$ . Тогда в этой норме выполняется оценка

$$\|\Pi u(t^n, \cdot) - u^n\| \leq c(K + 1) \left( h^P \left\| \frac{\partial^P u}{\partial x^P} \right\| + t^n h^Q \left\| \frac{\partial^{Q+1} u}{\partial x^{Q+1}} \right\| \right), \quad (18)$$

где  $n \in \mathbb{N}$  – номер шага по времени,  $u(t^n, \cdot)$  – временное сечение функции  $u(t, x)$  на время  $t^n$ ,  $u^n$  – численное решение по схеме (6) с начальными данными  $\Pi u(0, \cdot)$ ,  $Q \in \mathbb{N}$  и  $P \in \mathbb{N}$  получены в результате работы алгоритма, а  $c$  – константа, не зависящая от  $n$  и  $h$ .

Доказательство. Случай  $Q = P = \tilde{p} - 1$  тривиален, поэтому рассмотрим только случай  $Q \geq P = \tilde{p}$ . Как и оператор  $\Pi$ , оператор  $\tilde{\Pi}^Q$  не зависит от времени и действует только на временное сечение функции. Пусть  $u_{i,j}^n$  – решение с начальными данными  $\Pi u_0$ , а  $\tilde{u}_{i,j}^n$  – с начальными данными  $\tilde{\Pi}^Q u_0$ . Запишем

$$\|\Pi u(t^n, \cdot) - u^n\| \leq \|\Pi u(t^n, \cdot) - \tilde{\Pi}^Q u(t^n, \cdot)\| + \|\tilde{\Pi}^Q u(t^n, \cdot) - \tilde{u}^n\| + \|\tilde{u}^n - u^n\|.$$

Коэффициенты корректора  $h^q C^q$  имеют порядок  $h^q$ . Первое слагаемое в правой части есть разность между двумя проекциями точного решения на момент времени  $t^n$  и поэтому имеет порядок  $O(h^P)$ . Второе слагаемое есть ошибка решения в смысле проектора  $\tilde{\Pi}^Q$ . Поскольку схема точна на полиномах порядка  $Q$ , она обладает пространственной аппроксимационной ошибкой  $O(h^Q)$ . Ошибка интегрирования по времени по условию имеет тот же порядок или выше. Таким образом, ошибка решения оценивается как  $O(Kt^n h^Q)$ . Последнее слагаемое есть разность между решениями с двумя разными начальными данными, различающимися на величину порядка  $O(h^P)$ , и поэтому в силу устойчивости имеет порядок  $O(Kh^P)$ . Отсюда следует искомая оценка.

Таким образом, найденные в результате алгоритма значения  $P$  и  $Q$  действительно являются формальным порядком точности и порядком точности в длительном счёте. При помощи метода нестационарного корректора [13] можно найти вид главных членов ошибки решения по схеме (6) и, в частности, таким образом доказать неулучшаемость оценки (18). Мы не будем на этом останавливаться.

Для схем вида (6) возможны следующие альтернативы.

1.  $Q = P = \tilde{p} - 1$ . В этом случае  $\|Pu(t^n, \cdot) - u^n\| = O(h^{\tilde{p}-1}t^n)$ , то есть порядок точности совпадает с порядком аппроксимации. Этот случай имеет место для однородных конечно-разностных схем на равномерных сетках.
2.  $Q = P = \tilde{p}$ . В этом случае  $\|Pu(t^n, \cdot) - u^n\| = O(h^{\tilde{p}-1}t^n)$ , то есть порядок точности на 1 превышает порядок аппроксимации. Этот случай характерен для конечно-разностных и конечно-объёмных схем на одномерных неравномерных сетках, а также конечно-элементных схем на равномерных и одномерных неравномерных сетках. Большинство схем на неравномерных сетках можно записать в рассматриваемом виде (6) с  $h = const$ , если ограничиться рассмотрением периодических сеток и блочной стратегией их измельчения (см. [13]).
3.  $Q > P = \tilde{p}$ . В этом случае численная ошибка имеет вид (18), то есть порядок точности на 1 превышает порядок аппроксимации и, кроме того, наблюдается повышенная точность при длительном счёте. Ниже мы установим этот факт для разрывного метода Галёркина на равномерной сетке.

Из соображений размерности (в силу уравнения (1) единица времени совпадает с единицей длины) члены в численной ошибке могут иметь вид  $h^{n+\delta}t^{1-\delta}\|\partial^{n+1}u/\partial x^{n+1}\|$ ,  $0 \leq \delta \leq 1$ . При этом в общем случае  $\delta$  может быть отлично как от 0, так и от 1. Например, для простейшей конечно-объёмной схемы (совпадающей с методом Галёркина с кусочно-постоянными базисными функциями) на неструктурированной сетке старший член имеет такой вид с  $n = 0$  и  $\delta = 1/2$  (см. [2, 14, 15]), а для схемы EBR3 на нерегулярных структурированных сетках некоторого специального вида в [16] для старшего члена установлено  $n = 1$  и  $\delta = 1/4$ . Настоящее исследование показывает, что для схем вида (6) при  $h = const$  случай нецелых  $\delta$  невозможен.

## Примеры

В качестве примера рассмотрим метод Галёркина на основе базисных функций 0-го, 1-го и 2-го порядков. Для  $p = 0$  имеем

$$M = \left\| \left\| 1 \right\| \right\|, \quad M\hat{L} = \left\| \left\| 0 \right\| \right\|.$$

Точность на линейной функции имеется:  $(Mf)^1 = 0$ , а на квадратичной функции – нет:

$$(Mf)^2 = \left\| \left\| 1/2 \right\| \right\| h^2$$

Система  $(hM\hat{L})C^2 = h^{-1}(Mf)^2$  несовместна, поэтому нестационарный корректор 2-го порядка растёт линейно со временем. Схема имеет 1-й порядок точности как формально, так и при длинном счёте.

Для  $p = 1$  предположим, что точки коллокации заданы как  $x_0 = 0, x_1 = 1$ . Имеем

$$M = \left\| \left\| \begin{array}{cc} 1/3 & 1/6 \\ 1/6 & 1/3 \end{array} \right\| \right\|, \quad M\hat{L} = \frac{1}{h} \left\| \left\| \begin{array}{cc} 1/2 & -1/2 \\ -1/2 & 1/2 \end{array} \right\| \right\|.$$

Как и в случае  $p = 0$ , точность на линейной функции имеется:  $(Mf)^1 = 0$ , а на квадратичной функции – нет:

$$(Mf)^2 = \left\| \left\| \begin{array}{c} -1/12 \\ 1/12 \end{array} \right\| \right\| h.$$

Система  $(hM\hat{L})C^2 = h^{-1}(Mf)^2$  совместна, поэтому имеет место сверхсходимость: порядок точности равен 2. Продолжим исследование для установления эффекта повышенной точности на длительном времени счёта. Решение, удовлетворяющее условию отбора, равно

$$C^2 = \left\| \left\| \begin{array}{c} -1/12 \\ 1/12 \end{array} \right\| \right\|.$$

Подставляя полученные значения в (15) для  $q = 2$ , получаем

$$(Mf)^3 = \left\| \left\| \begin{array}{c} -1/18 \\ 1/18 \end{array} \right\| \right\| h^2.$$

Система  $(hM\hat{L})C^3 = h^{-2}(Mf)^3$  снова совместна. Поэтому нестационарный корректор 3-го порядка имеет порядок  $O(h^3)$  и при  $t \rightarrow \infty$  выходит на предельное значение  $C^3 h^3$ , где  $C^3$  является решением этой системы:

$$C^3 = \left\| \left\| \begin{array}{c} -1/18 \\ 1/18 \end{array} \right\| \right\|.$$

Подставляя полученные значения в (15) для  $q = 3$ , получаем

$$(Mf)^4 = \left\| \begin{array}{c} -1/27 \\ 5/216 \end{array} \right\| h^3.$$

Система  $(hM\hat{L})C^4 = h^{-3}(Mf)^4$  несовместна, поэтому нестационарный корректор 4-го порядка растёт линейно со временем и, следовательно, имеет порядок  $O(h^3)$ . Поэтому при длинном счёте старший член ошибки решения имеет 3-й порядок малости по  $h$ .

Для  $p = 2$  предположим, что точки коллокации заданы как  $x_0 = 0, x_1 = 1/2, x_2 = 1$ . Имеем

$$M = \left\| \begin{array}{ccc} 2/15 & 1/15 & -1/30 \\ 1/15 & 8/15 & 1/15 \\ -1/30 & 1/15 & 2/15 \end{array} \right\|, \quad M\hat{L} = \frac{1}{h} \left\| \begin{array}{ccc} 1/2 & 2/3 & -7/6 \\ -2/3 & 0 & 2/3 \\ 1/6 & -2/3 & 1/2 \end{array} \right\|,$$

$$(Mf)^3 = \left\| \begin{array}{c} 1/180 \\ -1/90 \\ 1/180 \end{array} \right\| h^2, \quad C^3 = \left\| \begin{array}{c} 1/60 \\ -1/240 \\ 0 \end{array} \right\|,$$

$$(Mf)^4 = \left\| \begin{array}{c} 7/1600 \\ -1/150 \\ 11/4800 \end{array} \right\| h^3, \quad C^4 = \left\| \begin{array}{c} 43/4800 \\ -19/9600 \\ -1/960 \end{array} \right\|,$$

$$(Mf)^5 = \left\| \begin{array}{c} 43/24000 \\ -7/3000 \\ 13/24000 \end{array} \right\| h^4, \quad C^5 = \left\| \begin{array}{c} 23/8000 \\ -9/16000 \\ -1/1600 \end{array} \right\|,$$

$$(Mf)^6 = \left\| \begin{array}{c} 2059/4320000 \\ -133/270000 \\ 223/1440000 \end{array} \right\| h^5.$$

Система  $(hM\hat{L})C^6 = h^{-5}(Mf)^6$  несовместна, поэтому нестационарный корректор 6-го порядка растёт линейно со временем и, следовательно, имеет порядок  $O(h^5)$ . Поэтому при длинном счёте старший член ошибки решения имеет 5-й порядок малости по  $h$ .

Коэффициенты, аналогичные приведённым выше, нами получены программным образом для метода Галёркина произвольного наперёд заданного порядка точности с использованием рациональной арифметики, заложенной в пакет `gmp` [17].

## Квадратуры типа Гаусса

Для дальнейшего исследования нам понадобятся  $(p + 1)$ -точечные квадратурные формулы интегрирования порядка  $2p$ , один из узлов которой фиксирован на правом конце отрезка.

Рассмотрим отрезок  $[0,1]$ . Покажем, что существует квадратурная формула

$$\int_0^1 f(x)dx = \sum_{k=0}^p v_k f(x_k), \quad (19)$$

где  $0 \leq x_k < 1$  для  $k = 0, \dots, p - 1$  и  $x_p = 1$ , точная на полиномах порядка  $2p$ .

Для этого запишем

$$\int_0^1 f(x)dx = f(1) + \int_0^1 (f(x) - f(1))dx = f(1) + \int_0^1 \frac{f(x) - f(1)}{1 - x}(1 - x)dx.$$

Для последнего интеграла запишем  $p$ -точечную квадратуру Гаусса порядка  $2p - 1$  с весом  $p(x) = 1 - x$ . Обозначим её узлы через  $x_k$ ,  $k = 0, \dots, p - 1$  и веса через  $w_k$ . Существование такой квадратуры, причём  $w_k > 0$  и  $0 < x_k < 1$ , доказано в [18]. Узлы этой квадратуры являются нулями многочлена Якоби с параметрами  $\alpha = 1, \beta = 0$  порядка  $p$ . Эти многочлены определяются рекуррентной формулой [19] (приводим перенормированную на отрезок  $[0, 1]$  формулу)

$$P_0 = 1, \quad P_1(x) = 3x - 1, \quad P_{n+1}(x) = (A_n(2x - 1) + B_n)P_n(x) - C_n P_{n-1}(x),$$

$$A_n = 2 - \frac{1}{n + 2}, \quad B_n = \frac{1}{(n + 2)(2n + 1)}, \quad C_n = 1 - \frac{2n + 2}{(n + 2)(2n + 1)}.$$

Имеем

$$\begin{aligned} \int_0^1 f(x)dx &\approx f(1) + \sum_{k=0}^{p-1} w_k \frac{f(x_k) - f(1)}{1 - x_k} = \\ &= \sum_{k=0}^{p-1} \frac{w_k}{1 - x_k} f(x_k) + \left(1 - \sum_{l=0}^{p-1} \frac{w_l}{1 - x_l}\right) f(1). \end{aligned}$$

Полученная квадратура точна на полиномах порядка  $2p$  и имеет искомый вид (19) с весами

$$v_k = \frac{w_k}{1 - x_k}, \quad k < p; \quad v_p = 1 - \sum_{j=0}^{p-1} v_j.$$

## Метод Галёркина с точками коллокации в узлах квадратуры

Всюду в настоящем разделе будем предполагать следующее.

1. Рассматривается метод Галёркина с разрывными базисными функциями (4) на основе многочленов порядка  $p$ .
2. Используется коллокационный базис, причём точки коллокации  $\{x_j\}$  выбраны в узлах полученной выше квадратуры, то есть  $x_j, j = 0, \dots, p-1$ , являются нулями полинома Якоби с  $\alpha = 1$  и  $\beta = 0$ , а  $x_p = 1$ .
3. Используется оператор проектирования  $\Pi$ , определённый формулой (12), то есть сопоставляющий произвольной функции  $f$  её значения в точках коллокации.

В силу предположения 2

$$\begin{aligned}
 m_{jk} &= \int_0^1 \phi_j(x)\phi_k(x)dx = \sum_{n=0}^p v_n \phi_j(x_n)\phi_k(x_n) = \begin{cases} v_j, & j = k; \\ 0, & j \neq k; \end{cases} \\
 g_{jk} &= \int_0^1 \phi'_j(x)\phi_k(x)dx = \sum_{n=0}^p v_n \phi'_j(x_n)\phi_k(x_n) = v_k \phi'_j(x_k); \\
 a_{jk} &= \delta_{jp}\delta_{kp} - g_{jk} = \delta_{jp}\delta_{kp} - v_k \phi'_j(x_k); \\
 b_{jk} &= -\delta_{j0}\delta_{kp}.
 \end{aligned} \tag{20}$$

**Утверждение 4.** В предположениях 1–3 справедливо  $(Mf)^{p+1} = 0$ , то есть имеет место  $(p+1)$ -й порядок аппроксимации.

Поскольку точность на полиномах порядка  $p$  имеется по построению, для доказательства этого утверждения найдём ошибку аппроксимации на функции  $-(x-t)^{p+1}/(p+1)!$ . Используя выражения (20) для  $m_{jk}$ ,  $a_{jk}$  и  $b_{jk}$ , получаем

$$\begin{aligned}
 h^{-p}(Mf)_j^{p+1} &= v_j \frac{x_j^p}{p!} - \sum_{k=0}^p v_k \phi'_j(x_k) \frac{x_k^{p+1}}{(p+1)!} + \frac{\delta_{jp}}{(p+1)!} = \\
 &= v_j \frac{x_j^p}{p!} - \int_0^1 \phi'_j(x) \frac{x^{p+1}}{(p+1)!} dx + \frac{\delta_{jp}}{(p+1)!} = \\
 &= v_j \frac{x_j^p}{p!} - \left( \frac{\delta_{jp}}{(p+1)!} - \int_0^1 \phi_j(x) \frac{x^p}{p!} dx \right) + \frac{\delta_{jp}}{(p+1)!} = 0.
 \end{aligned}$$

Утверждение доказано.



Проведём анализ разностной схемы (4) при помощи описанного выше алгоритма. Пусть выполнение алгоритма находится на некотором шаге  $q$ , следовательно, в ходе выполнения алгоритма найдены все коэффициенты  $C_j^m$ ,  $m \leq q$ ,  $j = 0, \dots, p$ .

Ключевым местом алгоритма на шаге  $q$  является анализ совместности системы (16) для нахождения коэффициентов  $C_j^{q+1}$ . Покажем, что при  $q \leq 2p$  эта система является совместной.

**Утверждение 5.** В предположениях 1–3 условие совместности системы (16) записывается в виде

$$\sum_{j=0}^p (Mf)_j^{q+1} = 0.$$

Доказательство. Матрица  $hM\hat{L} = A + B$  имеет коэффициенты

$$(m\hat{l})_{jk} = -v_k \phi'_j(x_k) + \begin{cases} 1, & j = k = p; \\ -1, & j = 0, k = p; \\ 0, & \text{otherwise.} \end{cases}$$

Матрица  $M\hat{L}$  имеет левый собственный вектор  $(1, \dots, 1)$ , соответствующий нулевому собственному значению. Поскольку функции  $\phi_j(x)$  линейно независимы, то из функций  $\phi'_j(x)$  можно составить единственную нетривиальную линейную комбинацию, тождественно равную нулю. Следовательно, других левых собственных векторов, соответствующих нулевому собственному значению, нет. Следовательно, условие совместности системы есть условие ортогональности её правой части строке  $(1, \dots, 1)$ , что и требовалось доказать.

Учитывая, что

$$\sum_j g_{jk} = \int_0^1 \frac{d}{dx} \left( \sum_{j=0}^p \phi_j(x) \right) \phi_k(x) dx = \int_0^1 \frac{d}{dx} (1) \phi_k(x) dx = 0,$$

суммируя (15) по  $j$ , имеем

$$\frac{1}{h^q} \sum_{j=0}^p (Mf)_j^{q+1} = \sum_{j=0}^p v_j \left( \frac{x_j^q}{q!} + \sum_{i=1}^q C_j^{q+1-i} \frac{x_j^{i-1}}{(i-1)!} \right) - \left( \frac{1}{(q+1)!} + \sum_{i=1}^q C_p^{q+1-i} \frac{1}{i!} \right).$$

Введём обозначение  $r^{q,m}$  для всех  $0 \leq m \leq q$  формулой

$$r^{q,m} = \sum_{j=0}^p v_j C_j^m \frac{x_j^{q-m}}{(q-m)!} - C_p^m \frac{1}{(q+1-m)!}. \quad (21)$$

В силу утверждения 4 при  $m \leq p + 1$  справедливо  $r^{q,m} = 0$ .

Поскольку  $\sum v_j x_j^q$  есть квадратурная формула для интегрирования  $x^q$  по отрезку  $[0, 1]$ , точная при  $q \leq 2p$ , справедливо  $\sum v_j x_j^q / q! = 1 / (q + 1)!$ . Отсюда

$$\frac{1}{h^q} \sum_{j=0}^p (Mf)_j^{q+1} = \sum_{i=1}^q r^{q,q+1-i}, \quad q \leq 2p. \quad (22)$$

**Утверждение 6.** В предположениях 1–3 при  $q \leq 2p$  справедливо

$$r^{q,m} = 0, \quad 0 < m \leq q. \quad (23)$$

Доказательство. Чтобы установить (23), рассмотрим систему уравнений (16) с правой частью (15) на шаге  $q - \alpha$ ,  $\alpha > 0$ , и просуммируем её по  $j$  с весом  $x_j^\alpha / \alpha!$ . Запишем вспомогательные равенства:

$$\sum_{j=0}^p m_{jk} \frac{x_j^\alpha}{\alpha!} = v_k \frac{x_k^\alpha}{\alpha!},$$

$$\begin{aligned} \sum_{j=0}^p g_{jk} \frac{x_j^\alpha}{\alpha!} &= \sum_{j=0}^p \int_0^1 \phi_j'(x) \phi_k(x) dx \frac{x_j^\alpha}{\alpha!} = \int_0^1 \frac{d}{dx} \left( \sum_{j=0}^p \frac{x_j^\alpha}{\alpha!} \phi_j(x) \right) \phi_k(x) dx = \\ &= \int_0^1 \frac{d}{dx} \left( \frac{x^\alpha}{\alpha!} \right) \phi_k(x) dx = v_k \frac{x_k^{\alpha-1}}{(\alpha-1)!}, \\ \sum_{j=0}^p a_{jk} \frac{x_j^\alpha}{\alpha!} &= -v_k \frac{x_k^{\alpha-1}}{(\alpha-1)!} + \delta_{kp} \frac{1}{\alpha!}, \end{aligned} \quad (24)$$

$$\sum_{j=0}^p b_{jk} \frac{x_j^\alpha}{\alpha!} = 0. \quad (25)$$

Система (16) для порядка  $q - \alpha$  записывается в виде

$$\sum_{k=0}^p (a_{jk} + b_{jk}) C_k^{q+1-\alpha} = \frac{1}{h^{q-\alpha}} (Mf)_j^{q+1-\alpha}. \quad (26)$$

Домножим  $j$ -е уравнение (26) на  $x_j^\alpha / (\alpha!)$  и сложим. С использованием (24) и (25) получаем

$$-\sum_{k=0}^p v_k C_k^{q+1-\alpha} \frac{x_k^{\alpha-1}}{(\alpha-1)!} + \frac{C_p^{q+1-\alpha}}{\alpha!} = \sum_{j=0}^p \frac{x_j^\alpha}{\alpha!} \frac{1}{h^{q-\alpha}} (Mf)_j^{q+1-\alpha}.$$

Левая часть этого равенства совпадает с определением (21) для  $m = q + 1 - \alpha$ . Подставляя правую часть из (15), получаем

$$\begin{aligned} r^{q,q+1-\alpha} &= \sum_{k=0}^p v_k \frac{x_k^\alpha}{\alpha!} \left( \frac{x_k^{q-\alpha}}{(q-\alpha)!} + \sum_{i=1}^{q-\alpha} C_k^{q+1-i-\alpha} \frac{x_k^{i-1}}{(i-1)!} \right) - \\ &- \sum_{k=0}^p \left( -v_k \frac{x_k^{\alpha-1}}{(\alpha-1)!} + \delta_{kp} \frac{1}{\alpha!} \right) \left( \frac{x_k^{q+1-\alpha}}{(q+1-\alpha)!} + \sum_{i=1}^{q-\alpha} C_k^{q+1-i-\alpha} \frac{x_k^i}{i!} \right). \end{aligned}$$

Группируя слагаемые без коэффициентов корректора и при коэффициентах корректора одного порядка, получаем

$$\begin{aligned} r^{q,q+1-\alpha} &= \left[ \sum_{k=0}^p v_k x_k^q \left( \frac{1}{\alpha!(q-\alpha)!} + \frac{1}{(\alpha-1)!(q+1-\alpha)!} \right) - \frac{1}{\alpha!(q+1-\alpha)!} \right] + \\ &+ \sum_{i=1}^{q-\alpha} \left[ \sum_{k=0}^p v_k C_k^{q-(i+\alpha-1)} x_k^{i+\alpha-1} \left( \frac{1}{\alpha!(i-1)!} + \frac{1}{(\alpha-1)!i!} \right) - C_p^{q-(i+\alpha-1)} \frac{1}{\alpha!i!} \right]. \end{aligned}$$

Вынесем за скобки общий множитель:

$$\begin{aligned} r^{q,q+1-\alpha} &= \frac{1}{\alpha!(q+1-\alpha)!} \left[ \sum_{k=0}^p v_k x_k^q (q+1) - 1 \right] + \\ &+ \sum_{i=1}^{q-\alpha} \frac{(i+\alpha)!}{\alpha!i!} \left[ \sum_{k=0}^p v_k C_k^{q+1-i-\alpha} \frac{x_k^{i+\alpha-1}}{(i+\alpha-1)!} - C_p^{q+1-i-\alpha} \frac{1}{(i+\alpha)!} \right]. \end{aligned}$$

Выражение  $\sum v_k x_k^q$  есть квадратурная формула для интегрирования функции  $x^q$  по отрезку. При  $q \leq 2p$  она точна, и сумма равна  $1/(q+1)$ . Поэтому выражение в квадратных скобках в первой строке равно 0. Во второй строке выражение в скобках совпадает с введённым ранее обозначением (21) для  $m = q + 1 - i - \alpha$ . Таким образом,

$$r^{q,q+1-\alpha} = - \sum_{i=1}^{q-\alpha} \frac{(i+\alpha)!}{\alpha!i!} r^{q,q+1-\alpha-i}, \quad (27)$$

Поскольку при  $q-\alpha \leq p$  коэффициенты  $C_j^{q-\alpha}$  равны 0, то и  $r^{q,q-\alpha} = 0$ ,  $q-\alpha \leq p$ . Из (27) по индукции по  $(q-\alpha)$  следует требуемое равенство (23):  $r^{q,m} = 0$  для всех  $m \leq q$ .

**Утверждение 7.** В предположениях 1–3 система (16) разрешима для  $q \leq 2p$ .

Следует из утверждения 5, формулы (22) и утверждения 6.

Комбинируя утверждения 3, 4 и 7, получаем следующий факт.

**Утверждение 8.** Пусть точки коллокации  $\{x_j\}$ ,  $j = 0, \dots, p-1$  выбираются в нулях полинома Якоби,  $x_p = 1$ , а оператор  $\Pi$  определён формулой (12). Тогда при  $p > 0$  справедлива оценка (18) с  $P = p + 2$  и  $Q = 2p + 1$ .

**Замечание 2.** При выполнении предположений 1–3 также наблюдается сверхсходимоссть в точке  $x = 1$ . А именно, выполняется равенство  $C_p^q = 0$  при  $q \leq 2p$ . Однако  $C_p^{2p+1} \neq 0$ . Этот факт не устанавливается приведённым доказательством и выявился при численном нахождении коэффициентов. Вероятно, он связан с аналогичным эффектом, имеющим место при интегральном операторе проектирования, см. [5].

## Случай неравномерной сетки

При описании алгоритма было отмечено (см. замечание 1), что условие ортогональности коэффициентов  $C^{q+1}$  левому собственному вектору  $hM\hat{L}$ , соответствующему нулевому собственному значению, не является обязательным. Вместо этого для каждого набора  $C^q$ ,  $q = p + 1, \dots, 2p + 1$ , наложим условие равенства нулю коэффициента в правой точке:  $C_p^q = 0$ .

Рассмотрим теперь неравномерную сетку с узлами  $X_i$ ,  $i \in \mathbb{N}$ . Определим  $h_i = X_{i+1} - X_i$ . Пусть точки коллокации имеют координаты  $x_{i,j} = X_i + x_j h_i$ , где  $x_j$  – узлы гауссовой квадратуры. Для анализа точности разрывного метода Галёркина на этой сетке рассмотрим оператор проектирования  $\tilde{\Pi}^{2p+1}$ , определяемый равенством

$$(\tilde{\Pi}^{2p+1}u)_{i,j}^n = u(t^n, x_{i,j}) + \sum_{q=p+2}^{2p+1} h_i^q C_j^q \frac{\partial^q u}{\partial x^q}(t^n, x_{i,j}). \quad (28)$$

Шаблон для вычисления  $du_{i,j}/dt$  разрывным методом Галёркина содержит только значения  $u_{i,k}$ ,  $k = 0, \dots, p$ , и  $u_{i-1,p}$ . Таким образом, аппроксимационная ошибка в смысле (28) в пределах ячейки  $i$  может зависеть от других ячеек только посредством множителя  $h_{i-1}^q C_p^q$ . А поскольку  $C_p^q = 0$ , этой зависимости не будет, и точность на полиномах порядка  $2p + 1$  в смысле (28) напрямую переносится с равномерной на произвольную неравномерную сетку. Полностью повторяя доказательство утверждения (3), получаем следующее.

**Утверждение 9.** Пусть  $u^n$  – решение на момент времени  $t^n$ , полученное разрывным методом Галёркина на неравномерной сетке, причём используется

точечный оператор проектирования при введённых выше точках коллокации. Пусть  $h = \max\{h_i\}$ . Тогда в условиях утверждения 3 справедлива оценка

$$\|\Pi u(t^n, \cdot) - u^n\| \leq c(K + 1) \left( h^{p+2} \left\| \frac{\partial^{p+2} u}{\partial x^{p+2}} \right\| + t^n h^{2p+1} \left\| \frac{\partial^{2p+2} u}{\partial x^{2p+2}} \right\| \right).$$

Отметим, что утверждение 9 справедливо для произвольных неравномерных сеток, в том числе, если при измельчении  $\max\{h_i\} / \min\{h_i\} \rightarrow \infty$ . Зависимость от неравномерности сетки может присутствовать только в константе устойчивости  $K$ .

### Случай произвольного базиса

Рассмотрим метод Галёркина с разрывными базисными функциями, причём на каждой сеточной ячейке решение является многочленом порядка  $p$ .

В предыдущем разделе рассматривался случай выбора коллокационных базисных функций с фиксированным набором точек коллокации  $\{x_j\}$ ,  $j = 0, \dots, p$ , и точечного оператора проектирования (12). При этом было доказано, что члены ошибки решения, растущие со временем, имеют порядок малости  $2p + 1$  и выше. Такое поведение ошибки связано с существованием оператора проектирования  $\tilde{\Pi}^{2p+1}$  на сетку, в смысле которого схема имеет  $(2p + 1)$ -й порядок аппроксимации. Этот оператор проектирования задаётся формулой (28).

Теперь предположим, что проектирование решения на сетку выполняется другим оператором проектирования  $\Pi^{new}$ , точным на полиномах порядка  $p$ . Например, таковым может быть точечный оператор проектирования в точках коллокации другого базиса или  $L_2$ -проекция на пространство полиномов в пределах каждой ячейки. Рассмотрим вопрос о точности разрывного метода Галёркина в этом случае.

Поскольку точность разностной схемы не зависит от выбора базиса в пространстве полиномов на каждой ячейке, а только от выбора проектора  $\Pi^{new}$ , будем рассматривать описанный выше коллокационный базис. Поскольку в смысле оператора проектирования  $\tilde{\Pi}^{2p+1}$  схема точна на полиномах порядка  $2p + 1$ , по аналогии с утверждением 3 получаем, что старший член численной ошибки, растущей со временем, имеет порядок  $h^{2p+1}$ . Отсюда следует, что  $Q = 2p + 1$ , то есть система (16) разрешима для  $q \leq 2p$ . Поскольку отсюда, в частности, следует, что система разрешима для  $q = p$ , получаем  $P \geq p + 1$ . Таким образом, выполняется оценка (18) с  $P \geq p + 1$  и  $Q = 2p + 1$ . Конкретное значение  $P$  определяется выбором оператора проектирования  $\Pi^{new}$ , при  $\Pi^{new} = \Pi$  выполняется  $P = p + 2$ .

Если используется другой базис, то схема будет точна на полиномах порядка  $2p + 1$  в смысле оператора проектирования  $\tilde{\Pi}'$ , действующего на функцию  $f$  следующим образом.

- Вычисляются значения в узлах гауссовой квадратуры по формуле (28).
- По полученным значениям строится лагранжев интерполяционный полином на каждой сеточной ячейке.
- Полином раскладывается по тому базису, который используется для построения схемы.

Если этот базис является коллокационным с другим выбором узлов  $\{x_j\}$ ,  $j = 0, \dots, p$ , то, взяв дифференциальное приближение от  $(\tilde{\Pi}' f)_j$  и отбросив старшие производные, можно привести его к виду (17) с другим набором коэффициентов. При этом система (16) разрешима для  $q \leq 2p$ , её условие совместности имеет тот же вид  $\sum (Mf)_j^{q+1} = 0$ . Однако каждый  $r^{q,m}$ , определённый (21), в отдельности может быть не равен 0; в этом можно убедиться, подставив приведённые выше значения коэффициентов  $C$  для  $p = 2$  при  $q = 4$ .

## Тестирование

Для верификации полученного результата рассмотрим уравнение (1) на отрезке  $[0,1]$  и функцию  $u(t, x) = \sin(2\pi(x - t))$ , являющуюся его решением. Последовательность расчётных сеток получим следующим образом: пусть первая сетка состоит из одного элемента – отрезка  $[0,1]$ , а каждая следующая сетка получается путём разбиения каждого элемента предыдущей на 3 части в пропорции  $0.2 : 0.3 : 0.5$ . Таким образом, максимальный размер сеточной ячейки  $n$ -й сетки равен  $0.5^n$ , а минимальный –  $0.2^n$ , причём максимальная и минимальная ячейки соседствуют между собой (с учётом периодических условий).

При расчёте по схеме на основе полинома порядка  $p$  будем использовать метод Рунге – Кутты  $(2p + 1)$ -го порядка аппроксимации для линейных задач, для уравнения  $du/dt = Au$  имеющий вид  $u^{n+1} = \left( \sum_{k=0}^{2p+1} A^k/k! \right) u^n$ . Чтобы избежать потери точности из-за ошибок округления, расчёт будем проводить с восьмерной точностью (quad double), заложенной в пакет QD [20]. Поскольку число шагов, требуемое для того, чтобы можно было пронаблюдать суперсходимость при больших временах, очень велико, на каждой сетке вычислялась матрица перехода со слоя на слой, а затем её нужная степень применялась к начальным данным.

Результаты для случая полиномов 4-го порядка ( $p = 4$ ) приведены на рис. 1. По горизонтальной оси отложено время, по вертикали – ошибка решения, вычисленная как максимальный по всем точкам коллокации модуль разности чис-

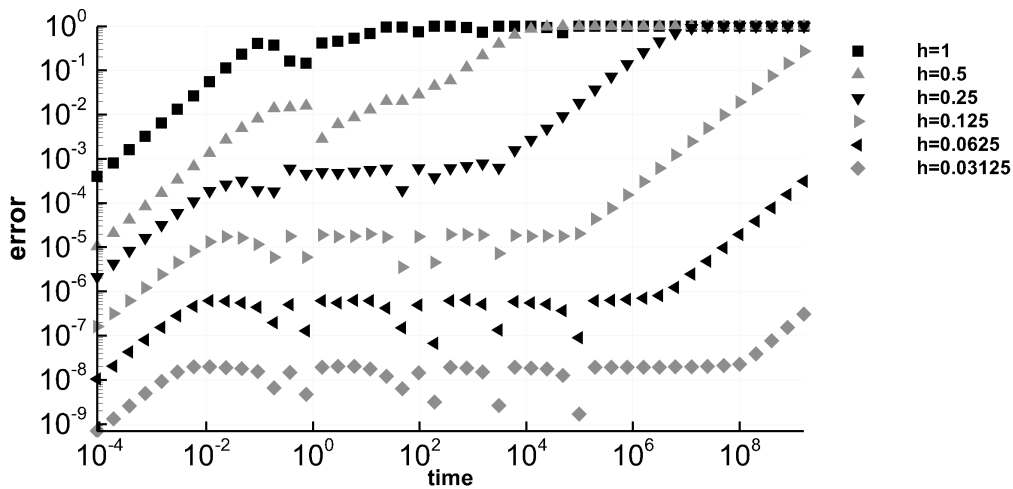


Рис. 1. Зависимость численной ошибки от времени в случае  $p = 4$

Таблица 1

**Численная ошибка для  $p = 4$  при  $t = 11.92$  и  $t = 10^8$**

$h_{\max}$	$e(t = 11.92)$	order	$e(t = 10^8)$	order
1	$6.81 \times 10^{-1}$	—	1	—
$2^{-1}$	$1.28 \times 10^{-2}$	5.73	1	—
$2^{-2}$	$5.59 \times 10^{-4}$	4.51	1	—
$2^{-3}$	$2.00 \times 10^{-5}$	4.80	$1.94 \times 10^{-2}$	—
$2^{-4}$	$6.16 \times 10^{-7}$	5.02	$1.95 \times 10^{-5}$	9.96
$2^{-5}$	$1.74 \times 10^{-8}$	5.14	$2.25 \times 10^{-8}$	9.75

ленного решения и проекции точного. Шкала на обеих осях логарифмическая. Результаты выведены на моменты времени  $t = 10^8/2^k$ ,  $k \in \mathbb{N}$ . Функция ошибки быстро осциллирует во времени, поэтому на графике видны “провалы”.

При  $t = 11.92$  и  $t = 10^8$  результаты сведены в таблицу 1.

## Заключение

В настоящей работе метод нестационарного корректора был применён к исследованию метода Галёркина с разрывными базисными функциями. Показано, что на равномерной сетке методом нестационарного корректора можно пользоваться «вслепую», то есть исследовать точность схемы, зная только её коэффициенты. Эта технология имеет естественное обобщение на многомерные трансляционно-симметрические сетки.

Для анализа систем вида (6) использованный метод требует только решения последовательности СЛАУ размера  $(p + 1) \times (p + 1)$  с известными коэф-

фициентами, определяемыми конкретной разностной схемой. Для сравнения, спектральный анализ, проведённый, в частности, в [12], требует нахождения собственных значений матрицы той же размерности, но зависящей от произвольного параметра  $\phi \in [0, 2\pi]$ . Это делает метод нестационарного корректора значительно более простым средством анализа точности разностных схем на равномерных и блочно-неравномерных сетках, которые могут быть представлены в виде (6). Недостатком метода нестационарного корректора является то, что он исходит из предположения устойчивости разностной схемы, тогда как спектральный анализ, напротив, устанавливает или опровергает устойчивость.

Также в настоящей работе был рассмотрен разрывный метода Галёркина на произвольной неравномерной сетке. В предположении устойчивости было доказано, что формальный порядок точности в одномерном случае равен  $p + 1$ , тогда как старший член в ошибке решения, неограниченно растущий со временем, имеет порядок  $h^{2p+1}$ .

## Список литературы

1. Reed W. H., Hill T.-R. Triangular mesh methods for the neutron transport equation: Tech. Rep.: 479: Los Alamos Scientific Laboratory, 1973.
2. Johnson C., Pitkaranta J. An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation // Mathematics of computation. 1986. Vol. 46. P. 1–26.
3. Cheng Y., Shu C.-W. Superconvergence of discontinuous Galerkin and local discontinuous Galerkin schemes for linear hyperbolic and convection-diffusion equations in one space dimension // SIAM J. Numer. Anal. 2010. Vol. 47. P. 4044–4072.
4. Shu C.-W. Superconvergence and long time evolution of DG method. Lecture notes in CEAA2012. 2012.
5. Enhanced accuracy by post-processing for finite element methods for hyperbolic equations / Cockburn B., Luskin M., Shu C.-W. et al. // Mathematics of Computation. 2002. Vol. 72. P. 577–606.
6. Zhang Q., Shu C.-W. Error estimates to smooth solutions of Runge – Kutta discontinuous Galerkin methods for symmetrizable systems of conservation laws // SIAM J. Numer. Anal. 2006. Vol. 44. P. 1703–1720.
7. Zhang Q., Shu C.-W. Stability analysis and a priori error estimates of the third order explicit Runge – Kutta discontinuous Galerkin method for scalar conservation laws // SIAM J. Numer. Anal. 2010. Vol. 48. P. 1038–1063.



8. Superconvergence of discontinuous Galerkin method for scalar nonlinear conservation laws in one space dimension / Meng X., Shu C.-W., Zhang Q. et al. // *SIAM J. Numer. Anal.* 2012. Vol. 50. P. 2336–2356.
9. Ryan J., Shu C.-W., Atkins H. Extension of a Post Processing Technique for the Discontinuous Galerkin Method for Hyperbolic Equations with Application to an Aeroacoustic Problem // *SIAM J. Sci. Comput.* 2005. Vol. 26. P. 821–843.
10. Postprocessing for the discontinuous Galerkin method over nonuniform meshes / Curtis S., Kirby R., Ryan J. et al. // *SIAM J. Sci. Comput.* 2007. Vol. 30. P. 272–289.
11. P. van Slingerland, Ryan J., Vuik C. Position-Dependent Smoothness-Increasing Accuracy-Conserving (SIAC) Filtering for Improving Discontinuous Galerkin Solutions // *SIAM J. Sci. Comput.* 2011. Vol. 33. P. 802–825.
12. Zhang M., Shu C.-W. An analysis of and a comparison between the discontinuous Galerkin and the spectral finite volume methods // *Computers and Fluids.* 2005. Vol. 34. P. 581–592.
13. Бахвалов П. А. Метод нестационарного корректора для анализа точности линейных разностных схем для уравнения переноса // *Препринты ИПМ им. М.В.Келдыша.* 2016. № 140. С. 1–32.
14. Peterson T. A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation // *SIAM J. Numer. Anal.* 1991. Vol. 28. P. 133–140.
15. Merlet B. Linf and L2-error estimates for a finite volume approximation of linear advection // *SIAM Journal on Numerical Analysis.* 2007. Vol. 46. P. 124–150.
16. Бахвалов П. А. О порядке точности рёберно-ориентированных схем на сетках специального вида // *Препринты ИПМ им. М.В.Келдыша.* 2017. № 79. С. 1–32.
17. *The GNU Multiple Precision Arithmetic Library.* URL: <https://gmplib.org/>.
18. Бахвалов Н. С., Жидков Н. П., Кобельков Г. М. Численные методы. М: Наука, 1987. с. 600.
19. Abramowitz M., Stegun I. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* New York: Dover, 1965. с. 561.
20. *QD – A C++/Fortran-90 double-double and quad-double package.* Yozo Hida, Xiaoye S. Li, David H. Bailey et al. URL: <https://github.com/aoki-t/QD>.